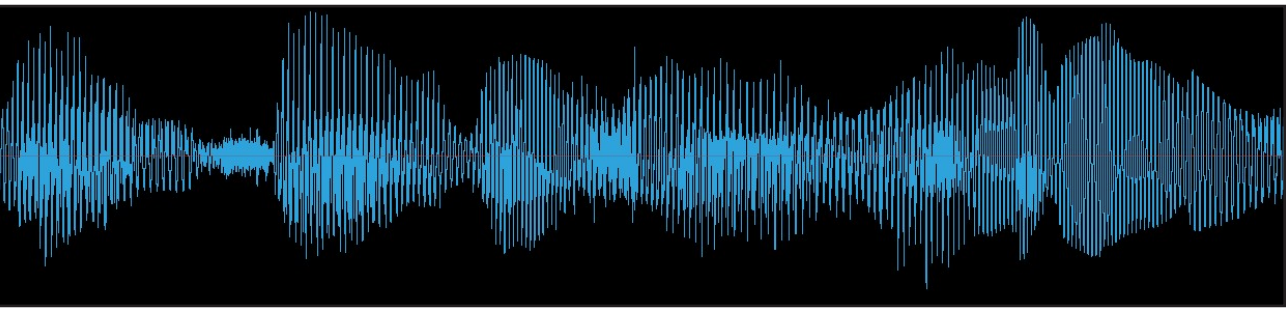
Pupillometry as a window to listening effort

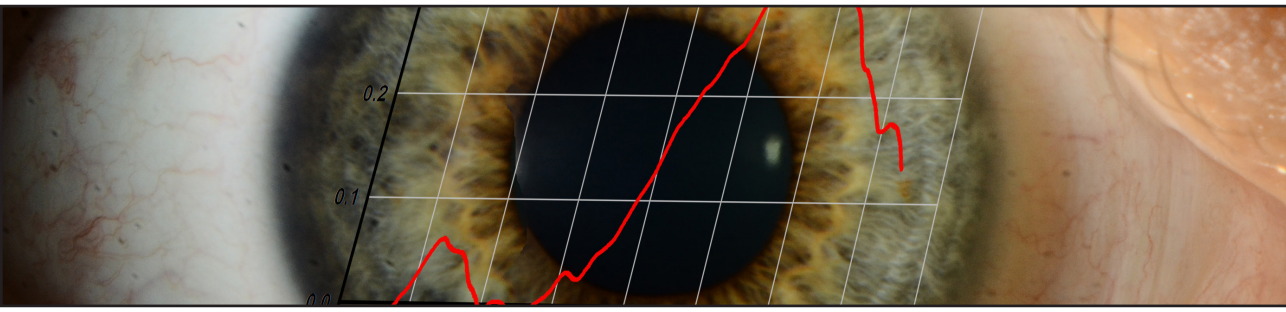
Ohlenforst, B.A.

2018

document version
Publisher's PDF, also known as Version of record



Chapter 3



Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation.

Ohlenforst, B.,
Zekveld, A. A.,
Lunner, T.,
Wendt, D.,
Naylor, G.,
Wang, Y.,
Versfeld, N. J.
Kramer, S. E.

Hearing Research (2017), 351, 68-79

Abstract

Previous research has reported effects of masker type and signal-to-noise ratio (SNR) on listening effort, as indicated by the peak pupil dilation (PPD) relative to baseline during speech recognition. At about 50% correct sentence recognition performance, increasing SNRs generally results in declining PPDs, indicating reduced effort. However, the decline in PPD over SNRs has been observed to be less pronounced for hearing-impaired (HI) compared to normal-hearing (NH) listeners. The presence of a competing talker during speech recognition generally resulted in larger PPDs as compared to the presence of a fluctuating or stationary background noise. The aim of the present study was to examine the interplay between hearing-status, a broad range of SNRs corresponding to sentence recognition performance varying from 0 to 100% correct, and different masker types (stationary noise and single-talker masker) on the PPD during speech perception. Twenty-five HI and 32 age-matched NH participants listened to sentences across a broad range of SNRs, masked with speech from a single talker (-25 dB to +15 dB SNR) or with stationary noise (–12 dB to +16 dB). Correct sentence recognition scores and pupil responses were recorded during stimulus presentation. With a stationary masker, NH listeners show maximum PPD across a relatively narrow range of low SNRs, while HI listeners show relatively large PPD across a wide range of ecological SNRs. With the single-talker masker, maximum PPD was observed in the mid-range of SNRs around 50% correct sentence recognition performance, while smaller PPDs were observed at lower and higher SNRs. Mixed-model ANOVAs revealed significant interactions between hearing-status and SNR on the PPD for both masker types. Our data show a different pattern of PPDs across SNRs between groups, which indicates that listening and the allocation of effort during listening in daily life environments may be different for NH and HI listeners.

3.1 Introduction

Hearing impairment has a variety of consequences, including reduced abilities to communicate, problems interpreting speech sounds, difficulties recognizing environmental sounds and even social isolation (Hirsh et al., 1952; Mathers et al., 2000). With impaired hearing, speech recognition becomes more challenging, especially when background noise is present (e.g. Arlinger, 2003; Dubno et al., 2015; Plomp, 1994; Plomp and Mimpen, 1979). This may lead to increased perceptual load (degree of selective attention processes for excluding distracting sensory information) and increased cognitive load (extent to which the task evoked demands consume available resources for successful task execution) compared to normal-hearing (NH) listeners (Pichora-Fuller et al., 2016). Even when hearing-impaired (HI) listeners' speech recognition performance is similar to that of NH listeners, the effort expended to accomplish the task is often greater for HI listeners (e.g. Fraser et al., 2010; Gatehouse and Gordon, 1990; Hällgren et al., 2005; Ohlenforst et al., 2017; Pichora-Fuller and Singh, 2006). Reduced fidelity of the auditory input signal results in higher need to invest mental effort to comprehend and respond appropriately to sound sources of interest. Listening effort has been defined as the deliberate allocation of mental resources to overcome obstacles to goal pursuit when carrying out a listening task (Pichora-Fuller et al., 2016). According to the framework for understanding effortful listening (FUEL) (Pichora-Fuller et al., 2016), listening effort depends not only on the individual's hearing ability, but also on the demands of the listening situation and the motivation of the listener to keep listening and not give up. The relationship between listening demand and success importance was recently demonstrated by measuring effort-related cardiovascular reactivity, an index of sympathetic activity, during an auditory discrimination task (Richter, 2016). Higher reward or success importance resulted in higher cardiovascular activity (higher sympathetic activity) (Richter, 2016). The interplay between listening demand and motivation is furthermore suggested by neuroimaging studies that indicate that supporting neural systems are adaptively applied during fruitful listening (Eckert et al., 2016). Listening is fruitful when the value of listening outweighs the relative costs of using these neural systems, for example when higher performance levels or rewards are obtained (Kouneiher et al., 2009) or when losses are avoided (Paulus et al., 2003). These findings support FUEL and indicate that the effort expended by a person during listening seems to be modulated by task demand and personal motivation to remain engaged in the task (Pichora-Fuller et al., 2016).

Commonly used intelligibility measures, such as word or sentence recognition, seem partly insensitive to different amounts of listening effort (Pichora-Fuller et al., 2016). For example, to maintain similar intelligibility levels during speech perception tasks, participants expended more mental effort in the presence of a single-talker masker than when stationary or fluctuating maskers were presented (Koelewijn et al., 2012). Recently, Wu and colleagues (2016) applied a dual-task paradigm with a primary sentence recognition task and two different secondary tasks (including either a simple visual reaction time task or an incongruent Stroop task) and additionally acquired subjective ratings to assess listening effort across a wide range of SNRs. They showed that as SNRs kept decreasing, speech recognition performance decreased. Surprisingly, reaction times became shorter (indicating

reduced effort) and subjective effort ratings were lower (indicating reduced effort) at the lowest SNRs (Wu et al., 2016). Listening may become so difficult that listeners decide to give up as the application of intense effort brings no further reward. At the highest SNRs, listening was very easy so that listeners did not need to expend much effort. In addition to the commonly applied speech perception tests, other measures that could provide more information about possible listening problems and effortful listening are required, and such measures need to be accessible at ecological SNR ranges (Lunner et al., 2016; Naylor, 2016).

Previous research has demonstrated that parameters derived from the task-evoked pupil responses, in particular the Peak Pupil Dilation (PPD), seem to reliably reflect listening effort, under various combinations of semantic or informational masking conditions and hearing abilities during speech recognition (Koelewijn et al., 2014, 2012; Zekveld et al., 2010; Zekveld and Kramer, 2014). One previous pupillometry study measured intelligibility conditions corresponding to sentence recognition performances between 0% correct and 99% correct in NH listeners with speech stimuli that were masked with interfering speech (Zekveld and Kramer, 2014). In line with Wu et al. (2016), the largest PPD resulted when about 50% correct sentence recognition was reached, relative to SNRs corresponding to lower and higher sentence recognition performance. The maximum PPD may differ for HI listeners as indicated by previous findings for speech recognition in stationary background noise. For example, the PPD during speech recognition in stationary background noise showed less decline with increasing SNR in HI compared to NH listeners (Zekveld et al., 2011). However, it is still unknown whether the pupil dilation for HI listeners differs from the pupil dilation for NH listeners when a single-talker masker is present. Recent research that investigated the SNRs hearing-aid users are exposed to (Smeds et al., 2015; Wu et al., 2016) demonstrated that hearing impaired listeners are exposed to a large range of SNRs in daily life sound environments. Positive SNRs ranging from +5 to +15 dB SNR cover the majority of daily life sound environments and communication situations. In line with this, an ecological momentary assessment of real-life situations showed that important communication situations included those in which HI listeners' speech intelligibility was typically rated as good or excellent but effort was rated as being high (Haverkamp et al., 2015). In contrast to these daily life conditions, speech recognition tests often include lower SNR ranges. As a result, our knowledge regarding the effort required in these ecological SNRs is limited, especially with respect to HI listeners. Therefore, in the present study, we aimed to examine how PPD varies across masker type and a broad range of SNRs that corresponds to performance scores across the entire psychometric function, in NH and HI listeners. We anticipated to learn more about the PPD at low, medium and high SNRs for the HI listeners and whether the PPD between NH and HI listeners would differ. We were furthermore seeking to answer whether the effect of masker type on the PPD would depend on the SNR or whether effects of masker type would show a consistent effect regardless of SNR. This study advances the findings by Wu and colleagues (2016) by the assessment of two essentially different masker types and the participation of age matched listener groups with different hearing abilities. We included a large range of positive SNRs to investigate how the pupil response changes when listeners are exposed to ecological listening situations. It is furthermore still not clear whether secondary task performance within dual-task paradigms (DTP) actually constitute an objective index of listening effort or 'mental exertion'. The multi-tasking paradigm

appears to have good validity to measure the ability to divide attention effectively in multi-tasking scenarios, but there is no independent way of measuring the resources dedicated to each task (McGarrigle et al,. 2014). A well-controlled pupillometry experiment on the other hand, can show relative task evoked changes in the pupil size which may reflect systematic changes in 'mental exertion' that cannot be obtained during behavioral measures alone (McGarrigle et al,. 2014).

We hypothesized low PPDs at very low and very high SNRs, either due to 'giving up' or due to 'very easy' conditions (Kramer et al., 1997; Zekveld et al., 2011; Zekveld and Kramer, 2014). We expected maximum PPD for SNRs in the mid-range at approximately 50% speech intelligibility (Zekveld and Kramer, 2014). We hypothesized that NH listeners would show larger PPD for difficult SNR conditions compared to HI listeners (see e. g. Koelewijn et al., 2014, 2012; Zekveld et al., 2011). This hypothesis was based on previous research showing less decline in PPD with increasing SNR for HI compared to NH listeners, as NH listeners had larger PPDs at more negative SNRs. HI listeners are more limited in their speech recognition performance as audibility and signal integrity can never be optimally restored, which may result in earlier performance surrender and consequently in a smaller pupil response (see e. g. Koelewijn et al., 2014, 2012; Zekveld et al., 2011). We furthermore hypothesized that the PPD changes would depend on masker type, with a larger PPD for sentence recognition in the presence of a single-talker masker compared to a stationary masker (Koelewijn et al., 2014, 2012). We expected that speech recognition in the presence of a single-talker masker would introduce more cognitive load and be more effortful due to informational masking, which would translate as a larger pupil response compared to speech recognition in the stationary noise masker (Koelewijn et al., 2014, 2012; Zekveld et al., 2014).

3.2 Method

Participants

The participants were recruited in the audiology clinic of the VU University Medical Center (VUMC) (HI group) and through flyers posted at the VUMC and around the VU University Campus (NH group). Both groups were age-matched within a range of five years, between 18 and 62 years old (mean age 47 years, SD=12.1) and native Dutch speakers. The audiometric inclusion criterion for the NH participants was a pure tone air conduction threshold average (PTA) <= 20 dB HL at 0.5, 1, 2 and 4 kHz in both ears. HI participants had symmetrical, mild to moderate sensorineural hearing thresholds with air-bone gaps less than 10 dB between 500 Hz and 4000 Hz in both ears, with PTAs for both ears between 35 dB and 60 dB HL. The pure-tone thresholds for those NH and HI participants included in the analysis, were averaged over both ears are presented in Figure 1. All participants had to have normal or corrected-to-normal vision, and no history of neurological diseases, dyslexia or diabetes mellitus. All participants provided written informed consent and the study was approved by the Ethics Committee of the VU University Medical Center in Amsterdam.

We calculated the effect size based on the mean PPDs provided by a recent pupillometry study in which two groups of listeners participated in a 50% correct sentence recognition test (Zekveld et al., 2011). A medium effect size of 0.3 was estimated using GPower, based

on the mean PPD from 38 middle-aged NH (mean PPD = 0.16 mm) and 36 middle-aged HI (mean PPD = 0.11 mm) listeners (Zekveld et al., 2011).
Power analysis revealed that a sample size of n=34 participants for each participant group (NH and HI) would provide a power of 0.79 to detect group (NH vs HI) effects with an expected effect size of 0.3 when applying an alpha-probability level of 0.05. We recruited and invited 35 NH and 35 HI participants for this study. Data from four HI participants were excluded due to unexpected changes in hearing thresholds with respect to an earlier audiogram (n=1), unexpected cognitive problems (n=1) or other health problems (n=2). The test session of one NH participant was not completed due to unfulfilled inclusion criteria regarding hearing status. Hence, data from 31 HI and 34 NH participants were included.

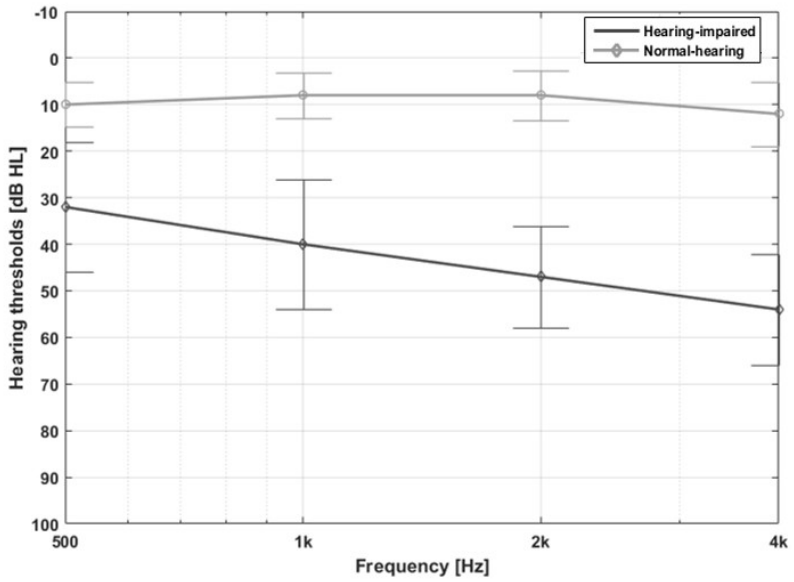


Figure1: Averaged pure tone hearing thresholds across 500 Hz, 1 kHz, 2kHz and 4 kHz for the hearing-impaired (HI) (dark-gray line) and the normal-hearing (NH) (light-gray line) participants. Error bars show the standard deviations of the mean.

Auditory stimuli

The auditory stimuli consisted of everyday Dutch sentences (Versfeld et al., 2000), which were presented binaurally via headphones. Sentences were similar to the HINT-sentences of Nilsson et al. (1994) and an extension to the sentence materials of Plomp & Mimpen (1979). Target sentences were spoken by a female talker. The number of syllables in each sentence was equal to 8 or 9. Words did not contain more than three syllables, and punctuation characters were absent. Sentences were mainly (92%) simple sentences, consisting of only one independent clause. An example sentence is: “de winkel is op loopafstand” (translation:

"the shop is within walking distance"). One percent of the sentences were compound (two independent clauses with coordinating conjunction), while 7% were complex, containing one independent and one dependent clause. Sentence duration ranged between 1.4 and 2 seconds. Speech recognition performance was measured in the presence of a stationary noise or a single-talker masker background. The single-talker masker consisted of concatenated sentences (Versfeld et al., 2000) spoken by a male voice. For each trial, the masker started 3 seconds before the presentation of the sentence, and ended 4 seconds after the sentence offset. An answer prompt tone was presented after the offset of the post-sentence noise, after which the participant repeated the sentence aloud. The answer prompt tone had a frequency of 1000 Hz, the presentation level was 55 dB SPL and the duration 1 second. The same presentation procedure was applied for both masker types. The experimenter scored the number of correctly repeated sentences. A sentence was scored as correct if the entire sentence was reproduced without mistakes. The long-term average frequency spectrum of both masker types was made identical to the spectrum of the target speech signal, and the masker was always presented at 65 dB SPL. The masker levels were kept constant to ensure that the noise would not become too loud at low SNRs. Keeping the masker level constant furthermore prevented listeners from learning to estimate task difficulty from changing noise levels, presented prior to the sentence in noise.

We aimed for SNRs that would provide performance scores across the whole psychometric function for each masker condition, including a large range of positive SNRs to cover ecological SNRs during daily life conditions. Previous research (Festen & Plomp, 1990), that measured psychometric functions for different masker types, including fluctuating noise, stationary noise and a single talker masker, showed that the speech reception thresholds (SRTs) for 50% correct sentence perception differed with up to 8 dB SNR between masker types. Furthermore, a difference of about 10 dB in SNR was shown between NH and HI listeners in the presence of a single-talker masker. In a previous study by Zekveld and Kramer (2014), young NH listeners performed a non-adaptive sentence recognition task in a single-talker masker condition across a wide range of negative SNRs. At about -25 dB SNR, the young NH listeners reached 0% correct sentence recognition performance. We assumed that HI listeners would also not be able to recall the target sentences at such low SNR conditions and we set the lowest SNR value to -25 dB for the single talker masker. Based on previous findings (Festen & Plomp, 1990; Zekveld & Kramer, 2014, Smeds et al., 2015; Wu et al., 2016) the SNR range for each masker type was chosen to estimate performance scores across the whole psychometric function of each masker type, including a larger range of positive SNRs to cover the majority of daily life sound environments and communication situations for HI listeners.

Speech masked with the stationary masker was presented at eight SNRs between -12 dB and +16 dB, distributed in steps of 4 dB. Speech masked with the single-talker masker was presented at nine SNRs between -25 dB and +15 dB, distributed in steps of 5 dB. Sentence recognition was measured in a randomized presentation order of SNRs for each masker type. Per masker type, ten sentences were presented for each SNR.

Spectral shaping

The HI participants did not wear hearing aids during the sentence recognition test. Instead, the sound files for the speech recognition tests were individually amplified according to the pure tone thresholds of each ear, by applying spectral shaping according to the NAL-R (Byrne and Dillon, 1986) prescription algorithm. It was applied in 1/3 octave steps within the frequency range of 315 Hz to 6300 Hz. The highest sound pressure level within each frequency band was set to 95 dB SPL to avoid headphone saturation. We verified audibility of the stimuli for all participants by testing sentence perception in quiet at 65 dB SPL. All participants reached 100% correct sentence recognition for 20 sentences presented in quiet.

Pupillometry

An eye tracking system by SensoMotoric Instruments (Berlin, Germany, 2D Video-Oculography, version 4), which applies infrared video tracking technology, was used to measure the pupil diameter during the experiment. The eye-tracking system had a spatial resolution of 0.03 mm and a sampling frequency of 60 Hz. During the experiment, the pupil location and the pupil size were recorded by the eye tracker and stored at a connected computer. The stored data included time stamps corresponding to the start of each trial, including the noise onset, the sentence onset, and the prompt tone, and the sentence recognition score, as entered by the experimenter. The experimenter monitored real-time pupil data during the experiment and applied corrective actions, such as the adjustment of the distance to the screen, or light adjustment, if needed.

Procedure

The test sessions were carried out in a sound proof booth and each participant sat on a fixed chair in front of a computer screen. The height of the chair and the distance to the screen (55 cm +/- 5 cm approx.) were adjusted individually until the conditions were optimized for the pupil recording. Each test session started with the calibration of the light conditions to avoid ceiling or floor effects in the pupil response (Hyönä et al., 1995). The pupil size was first measured during a condition of maximum illumination (230 lux) and afterwards in darkness. The illumination was individually adapted for each participant until the pupil size reached the middle of the dynamic range between maximum illumination and darkness. The mean illumination in the measurement booth was 13.3 lux (SD = 3.2 lux).
Each participant's visit started with a practice session to ensure confidence with the experimental procedure as it may not be intuitive to focus on a fixation dot and to inhibit movements and blinking during the sentence presentation. In this session, a single sentence for each SNR by masker type condition was randomly presented, resulting in 17 sentences in total. After the practice round, the sentence recognition test started with 20 sentences presented in quiet. Then the two experimental blocks were presented in random order. For each sentence, the pupil diameter was recorded. The participants were asked to focus on a white fixation dot on the blank computer screen in front of them and to inhibit eye blinks

during the presentation and response intervals. After each block (lasting between 12 and 15 minutes), the participants took a break of about 10 minutes. For each of the 17 conditions (8 SNRs for stationary masker, 9 SNRs for single-talker masker), pupil traces were recorded for 10 sentences per condition. In total, 170 sentences were presented per participant and one pupil trace was recorded per sentence.

Pupil data selection, cleaning and data reduction

The average pupil diameter recorded during the final second of the three second presentation of the masker, before target speech onset, was computed and used as baseline. The mean pupil diameter between the onset of the sentence and the answer prompt tone was calculated relative to the baseline pupil diameter for every trace (one pupil trace was recorded per sentence). The maximum pupil diameter between the onset of the sentence and the response prompt, relative to the baseline pupil diameter, is the peak pupil dilation (PPD). Pupil diameter values more than 3 standard deviations below the mean pupil diameter (between sentence onset and prompt tone relative to the baseline) were defined as blink. Traces with more than 15% of blinks between the start of the baseline (last second of pre-noise before sentence onset) and the prompt tone were excluded from the data analysis. Blinks were replaced by linear interpolation, starting 5 samples before and 7 samples after a blink (Siegle et al. 2008). The pupil response within each selected and de-blinked trace was smoothed by a 7-point moving average filter. For each participant, all the included de-blinked and smoothed traces (max. 10) for each condition were time-aligned and averaged. The PPD of this averaged trace provided the data for the statistical analysis.

Statistical analyses

Pupil data selection, cleaning and data reduction was applied to pupil data from 34 NH and 31 HI participants. For two of the 34 NH participants and six of the 31 HI participants, we identified less than 5 valid pupil traces out of the 10 pupil traces recorded per condition, across all conditions. Pupil and intelligibility data for those participants were consequently excluded and data for 32 adults with normal hearing (mean age 47.8 years) and for twenty-five adults with hearing impairment (mean age 47.9 years) were further analyzed. Five of 32 NH and six of 25 remaining HI participants had missing PPD values (more than 15% blinks across the 10 pupil traces per condition) across one (or more) of the 17 conditions. We measured 170 pupil traces per participant and on average, 15.8 (SD=16.2) pupil traces were missing per person. We applied linear mixed models (LMM) to analyze the data as LMM's tolerate missing values and do not exclude participants with missing values from the analyses. A linear mixed-effects model was built in R-studio using the packages *lme4* (Bates et al., 2015) and the function *lmer* was applied to fit LMM to the data. Two separate LMM ANOVAs were used to test the effect of SNR for each masker type (single-talker masker and stationary noise masker) on the PPD and percent-correct sentence recognition. The averaged PPD or percentage correct sentence recognition scores for each SNR were the dependent measures with participants as the repeated measure and therefore the random

effects. The fixed effects in each separate LMM ANOVA were the categorical variables group, SNR and the interaction between group and SNR. We did not include (random) interactions between SNR and participants as random factor in the separate 2-way LMM ANOVA's as our data did not include replicated data for combinations of SNR and participants. The setting is a classic randomized block setting with participants as blocks and SNR as treatments. The main effects of SNR and group (NH vs. HI) and the interaction between SNR and group were examined. To consider effect size estimates as supplement for the p-values for the output of our mixed models the so called "plug-in" method was applied where a back transformation of the F-statistics from a purely fixed-effect version of the models is used to compute delta-tilde (Brockhoff et al., 2016, section 4.4 to 4.5). The F-statistics itself is generally not the best measure of effect size as it depends on the number of observations for each product (Brockhoff et al., 2016). A delta-tilde, on the other hand, corresponds to an average of a number of Cohen's ds measuring the average pairwise effect size. Some of the effects had relatively large F-values, which produced large effect sizes and therefore large delta-tilde values. In general, larger delta-tilde values produce larger effect sizes and small delta-tilde values produce small effect sizes, as is the case for Cohen's d (Cohen, 1988).

A statistical comparison of both masker types requires that PPD data and %correct performance scores are available at the same SNR values. The stationary noise masker was presented across the SNR range between -12 to +16 in steps of 4 dB and the single-talker masker was presented from -25 dB to +15 dB SNR, in 5 dB steps. We selected an overlapping SNR range for both maskers from -10 dB to +15 dB SNR in 5 dB steps, which included part of the measured SNRs for the single-talker masker. For the stationary noise masker, psychometric functions were fitted to the individual performance scores of every participant across the originally measured SNRs (-12 to +16 dB SNR) to estimate the data points for the analyzed SNRs in the overlapping SNR range. The SNR and the slope at the individual 50% correct performance level was estimated, based on the computation of a logistic discrimination function across a range of SNRs (Green & Swets, 1966). The resulting psychometric functions covered the overlapping SNR range of -10, -5, 0, +5, +10, +15 dB SNR and the performance scores at those SNRs were individually estimated for every participant. The same principle was applied to estimate pupil data corresponding to the SNR range from -10 to +15 dB SNR for the stationary noise masker. Therefore, spline curves were fitted to the originally measured pupil data. A new data set was created with PPD and %correct sentence recognition performance for the overlapping SNR range from -10 dB to +15 dB for both masker types. For the statistical comparisons of the effect of masker types, the originally applied LMM ANOVA was extended by a fixed effect variable corresponding to the different masker types, by the 2-way interactions between masker and group, SNR and masker, SNR and group, and by the 3-way interaction between group, SNR and masker type. The participants were treated as repeated measures and included as random factor. A linear mixed-effects model was built in R-studio using the packages lme4 (Bates et al., 2015) and the function lmer was applied to fit LMM to the data, as done for the two separate LMM ANOVAs. The new dataset for the overlapping SNR range included replicated PPD data for the combination of SNR and masker type per participant. Two observations were made per SNR as two different noise types were tested at the same SNR values. To keep the random effects in the 3-way model maximal, the repeated observations required that the (random)

interactions between SNR and participant and between masker type and participant were added in the statistical model. We included the random effect of SNR as random slope of SNR, to allow each participant to have their own mean PPD size and their own effect of SNR on PPD with SNR nested within participants.
The lmerTest package offers the function rand, which allows to perform a likelihood ratio test on the random effects of our LMM (Kuznetsova et al., 2016). The Chi square statistics and the corresponding p-values of likelihood ratio tests are the output of the function rand and were reported for the analysis of the random effects. The function rand provides a likelihood ratio test that is the comparison of a model with a given random factor to that model without the random factor. The Chi square statistics indicate the variability in the outcome measures PPD and % correct sentence performance across participants, depending on the SNR and masker type. The effect sizes were estimated by applying the "plug-in" method as described for the separate LMMs above. The package lsmeans, including the function lsmeans was used to apply pairwise comparisons for post hoc analysis of significant interaction effects.

3.3 Results

Sentence recognition data

Figure 2 shows the sentence recognition scores across the range of SNRs for the stationary noise masker for NH and HI participants. Error bars represent the standard error of the mean. Both groups have 0 % sentence recognition at -12 dB SNR and reach close to 100 % correct sentence recognition between +8 and +16 dB SNR. A shift of about 2 dB was observed between the sentence recognition curves for the two groups at the middle range of SNRs, with poorer performances for the HI listeners. A mixed-model ANOVA revealed significant main effects of SNR ($F[7,371]=1379.4$, $p<0.001$, effect size delta-tilde=174,90) and group ($F[1,53]=46.6$, $p<0.001$, effect size delta-tilde=39.82), and a significant interaction between group and SNR ($F[7,371]=17.33$ $p<0.001$, effect size delta-tilde=1.38). After applying a Bonferroni correction to account for multiple testing across eight SNRs ($p=0.05/8$, i.e. 0.006), significant differences between the NH and HI group remained at SNRs of -4 dB, 0 dB and +4 dB (indicated by gray stars in Figure 2).

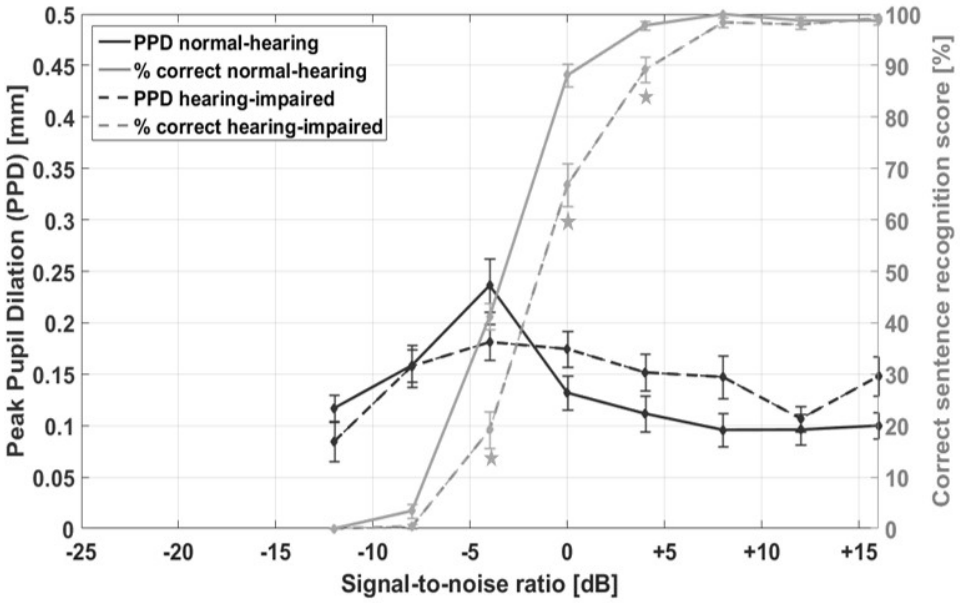


Figure2: Peak pupil dilation (PPD) (black color) on the left y-axis and percentage correct sentence recognition scores (gray color) on the right y-axis across signal-to-noise ratios (SNRs) for the stationary masker for normal-hearing (NH) and hearing-impaired (HI) participants. Error bars represent the standard error of the mean. Gray stars indicate significant group differences in sentence recognition performance (NH vs. HI) of $p<0.006$.

Figure 3 shows the sentence recognition scores across the range of SNRs for the single-talker masker for NH and HI participants. Error bars represent the standard error of the mean. Both groups have 0 % sentence recognition at -25 dB SNR and reach close to 100 % correct sentence recognition at +10 and +15 dB SNR. A shift of about 10 dB was observed between the sentence recognition curves for the two groups at the middle range of SNRs, with poorer performances for the HI listeners. A mixed-model ANOVA revealed significant main effects of SNR ($F[8,432]=551.7$, $p<0.001$, effect size delta-tilde=62.02) and group ($F[1,54]=106.5$, $p<0.001$, effect size delta-tilde=147.89), and a significant interaction between group and SNR ($F[8,432]=31.22$ $p<0.001$, effect size delta-tilde=1.89). After applying a Bonferroni correction to account for multiple testing across nine SNRs ($p=0.05/9$, i.e. 0.0055), significant differences between the NH and HI group remained at SNRs of -15 dB, -10 dB, -5 dB, 0 dB and +5 dB (indicated by gray stars in Figure 3).

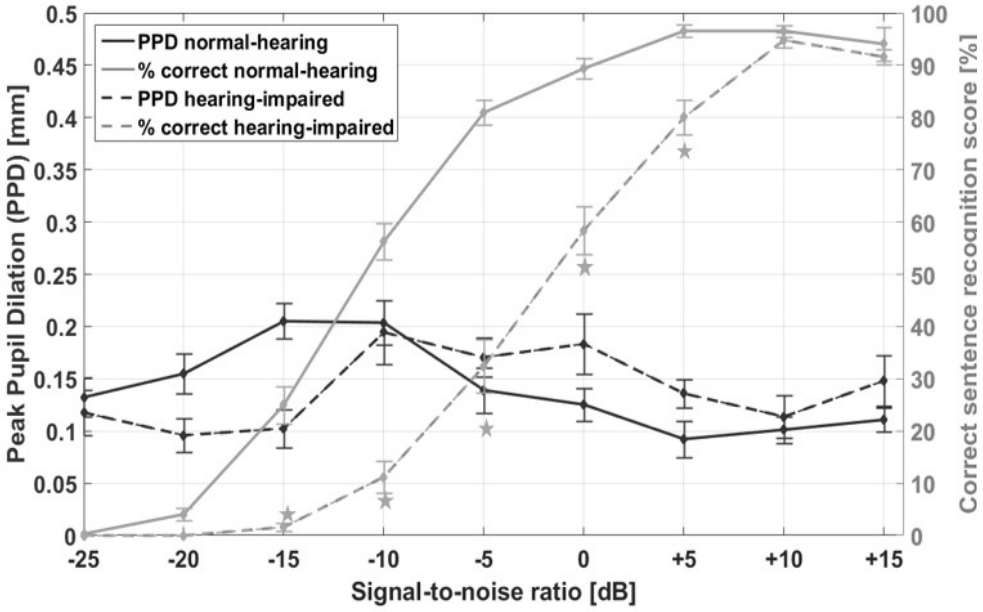


Figure3: Peak pupil dilation (PPD) on the left y-axis (black color) and percentage correct sentence recognition scores (gray color) on the right y-axis across signal-to-noise ratios (SNRs) for the single-talker masker for normal-hearing (NH) and hearing-impaired (HI) participants. Error bars represent the standard error of the mean. Gray stars indicate significant group differences in sentence recognition performance (NH vs. HI) of $p<0.0055$.

Figure 4 shows sentence recognition across the overlapping range of SNRs from -10 to +15 dB SNR for both masker types and both groups of listeners. Error bars represent the standard error of the mean. For the stationary noise masker (gray solid (NH) and gray dashed (HI) lines), both listener groups had similar performances (<10% correct) at the lowest SNR of -10 dB. When the SNR increased by 5 dB (at -5 dB SNR), NH listeners performed about 11 % better than HI listeners. The performance difference between groups decreased gradually with increasing SNRs until 100% correct sentence recognition performance was reached at +10 dB SNR. The difference between the groups in performance was larger for the single-talker masker (solid (NH) and dashed (HI) lines) as compared to stationary noise. At -10 dB SNR, NH listeners had about 41% better sentence recognition than HI listeners. This performance difference between groups decreased gradually until high performances (i.e. ~ 90% correct for HI and ~100% correct for NH listeners) were reached. We applied a LMM ANOVA on sentence recognition (percent correct) across SNRs for both masker types and both groups. The analysis revealed a significant main effect of SNR ($F[5,270]=936.34$, $p<0.0001$, effect size delta-tilde=139.77), a significant main effect of masker type ($F[1,54]=79.34$, $p=0.0001$, effect size delta-tilde=49.84) and a significant main effect of group ($F[1,54]=84.67$, $p<0.0001$, effect size delta-tilde=137.76) on sentence recognition. The three two-way interactions

(group x SNR, group x masker type and SNR x masker type) and the three-way interaction (group x SNR x masker type) were highly significant (all p-values < 0.0001). To analyze the random effects (participant, participant x masker type and participant x SNR) on sentence recognition performance, a Chi-square test was performed. The Chi-square statistics revealed that the random factor participant was significant ($\chi^2(1, N=54)= 13.16$, $p<0.0001$), indicating that sentence correct performance differed significantly across participants. The interaction between participant and masker type ($\chi^2(1, N=54)=5.66$, $p=0.017$) and the interaction between participant and SNR ($\chi^2(1, N=54)=9.02$, $p=0.003$) were also significant. The sentence recognition performances differed significantly between participants depending on the SNR and masker type.

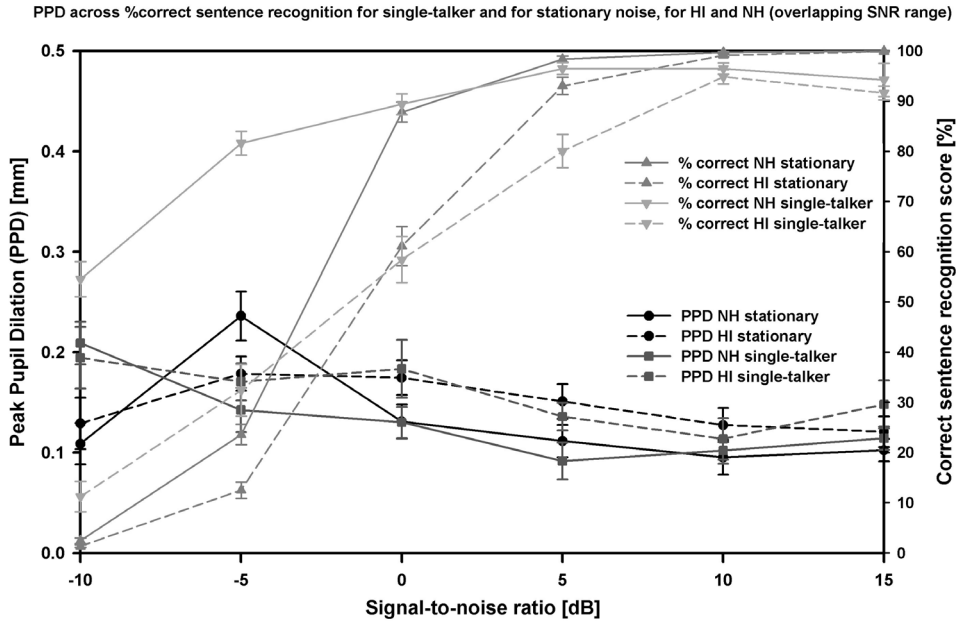


Figure4: Peak pupil dilation (PPD) (circles and squares) and percentage correct sentence recognition scores (triangles) across signal-to-noise ratios (SNRs) for the stationary and the single-talker masker for normal-hearing (NH) and hearing-impaired (HI) participants. The percentage correct sentence recognition scores were estimated based on the fitted psychometric function for the stationary noise masker and based on the measured sentence recognition scores for the single-talker masker. PPDs for were estimated based on curve fitting for the stationary noise masker and based on measured PPD for the single-talker masker. Error bars represent the standard error of the mean.

Based on the measured sentence recognition scores for the single-talker masker and the fitted psychometric functions for the stationary noise masker, the SNRs and the slopes at 50% correct were estimated (see Table 1). The SNRs at 50% correct were lower for the NH than for the HI listeners for both maskers. The estimated slopes of the psychometric

functions for stationary noise at 50% correct performance were slightly steeper for NH (13.7 %/dB) listeners than for HI listeners (11.2 %/dB). For the single-talker masker, the slopes at 50% correct sentence recognition performance were less steep than for the stationary noise masker but very similar for both groups of listeners (NH: 5.0 %/dB and HI: 5.6 %/dB). The steepness of the slopes for the stationary noise masker is smaller than previous findings, that showed a slope of 21.0% for a steady-state masker and 11.9% per dB for a two-band modulated noise masker when speech recognition was tested for normal-hearing listeners (e. g. Festen and Plomp, 1990).

Table1: Signal-to-noise ratio (SNR) and slope of the psychometric functions at 50% correct sentence recognition performance. SNR and slope values were estimated based on the fitted psychometric function for the stationary noise masker and based on the measured sentence recognition scores for the single-talker masker.

Listener group	Masker type	SNR [dB] at 50% correct	Slope [%/dB] of psychometric function at 50% correct
Hearing-impaired	Single-talker masker	-1.8	5.0
Normal-hearing	Single-talker masker	-11.2	5.6
Hearing-impaired	Stationary noise masker	-0.80	11.2
Normal-hearing	Stationary noise masker	-3.13	13.7

Pupil data

In Figure 2, the averaged PPD across SNRs is shown for the stationary noise masker for NH (solid, black line) and HI (dashed, black line) participants. The NH participants had a maximum PPD at -4 dB SNR, where they achieved about 42% correct sentence recognition (solid, gray line). HI listeners (dashed, black line) had relatively large PPDs across a wide range of SNRs, where they were achieving 60-100% correct sentence recognition. A mixed-model ANOVA on the PPD across SNRs for the stationary masker revealed a significant main effect of SNR ($F[7,356.85])=10.8$, $p<0.001$, effect size delta-tilde=2.18) but there was no significant main effect of listener group. The analysis also revealed a significant interaction effect between group (HI versus NH) and SNR ($F[7,356.85]=2.79$, $p=0.008$, effect size delta-tilde=0.40), indicating that the response across SNRs varied between listener groups. We built another model for the LMM ANOVA to further investigate the significant interaction between group and SNR. The model implied adjustments in the order of the fixed factors to

estimate differences in PPD across groups and SNRs. The analysis revealed that the PPD was not significantly different between groups but that the PPD differed across SNRs within each group. We applied a Bonferroni correction to account for multiple testing across eight SNRs ($p=0.05/8$) resulting in a p-value of 0.006. Within the NH group, PPD values corresponding to -8 ($t[354]=3.19$, $p=0.002$) and -4 dB ($t[354]=4.81$, $p<0.001$) SNR differed significantly from the lowest PPD value at -12 dB SNR. The PPD for the HI listeners differed significantly at -8 ($t[354]=3.19$, $p=0.002$), -4 ($t[354]=4.98$, $p<0.001$), 0 ($t[354]=4.44$, $p<0.001$), +4 ($t[354]=3.63$, $p<0.001$) and +16 dB ($t[354]=3.16$, $p=0.002$) SNR from the PPD at -12 dB SNR.

The average PPDs for the NH (solid, black line) and the HI (dashed, black line) participants in the single-talker masker conditions are shown in Figure 3. For the NH participants, maximum PPD was measured at -15 and -10 dB SNR (solid, black line), corresponding to 25% and 55% correct sentence recognition performance, respectively (gray, solid line). The HI participants showed their largest PPDs between -10 dB and 0 dB SNR, where they were achieving between 12% and 58% correct sentence recognition. A mixed-model ANOVA revealed a significant main effect of SNR ($F[8,395.14] = 5.27$, $p<0.001$, effect size delta-tilde=0.67) but no significant main effect of group on the PPD. The analysis also revealed a significant interaction between SNR and group (HI versus NH) ($F[8,395.14] = 6.56$, $p < 0.001$, effect size delta-tilde=0.80). As for the stationary noise masker, the interaction between group and SNR was further investigated by estimating group effects for the same SNRs. After applying a Bonferroni correction to account for multiple testing across nine SNRs ($p=0.05/9$, i.e. 0.0055), significant differences between the NH and HI group remained at -20 dB SNR ($t[390]=3.5$, $p<0.001$) and at -15 dB SNR ($t[390]=4.7$, $p<0.001$). Significantly larger PPDs were observed for the NH compared to the HI listeners. At positive SNRs, no significant effect of group on the PPD was found.

In Figure 4, the PPDs across the overlapping range of SNRs from -10 to +15 dB are shown for both masker types and both groups of listeners. The LMM ANOVA on the PPD across SNRs for both masker types revealed a significant main effect of SNR ($F[5,78.2]=8.76$, $p<0.0001$, effect size delta-tilde=2.18) and a significant interaction effect between SNR and masker type ($F[5,375.5]=9.28$, $p<0.0001$, effect size delta-tilde=0.40). The main effects of group and masker type, the two-way interactions between group x SNR, group x masker type and the three way interaction between group x SNR x masker type were not significant. To analyze the random effects (participant, participant x masker type and participant x SNR) on PPD, a chi-square test was performed. The random factor participant was significant ($X2(1, N=54)=11.56$, $p<0.0001$), indicating that the PPD differed significantly across participants. A significant interaction between participant and masker type ($X2(1, N=54)=6.57$, $p=0.01$) was found, indicating that changes in PPD across participants differed significantly depending on the masker type. No significant relationship between participant and SNR was found ($X2(1, N=54)=3.0$, $p=0.08$), indicating that there was no significant inter-individual variation in PPD per SNR.

The PPD values were averaged over the level of group and a Tukey post-hoc test on the significant interaction between SNR and masker type revealed differences within and between masker types across SNR. This method implies an adjustment of p-values for comparing a family of 12 estimates. The pairwise comparison in Table 2 shows that the PPD for the SNR at -10 dB in the single talker masker and the PPD at -5 dB SNR in the stationary noise masker were most different across SNRs and masker types. PPDs measured at -5 dB

SNR in stationary noise differed from the PPDs at +5, +10, and +15 dB SNR for the single-talker masker. The difference between the two masker types is not restricted to one SNR (-10 dB) only. For both masker types, the PPD measured at a low SNRs differed significantly from a range of higher, positive SNRs.

Table2: Resulting p-values for across masker-type comparison from Tukey post hoc test on the significant interaction effect between SNR and masker type. Results are averaged over levels of group and correction for multiple comparison was taken into account.

Masker type SNR [dB]	Single-talker -10 dB	Single-talker -5 dB	Single-talker 0 dB	Single-talker +5 dB	Single-talker +10 dB	Single-talker +15 dB
Stationary noise -10 dB	p<0.0001	p=0.61	p=0.61	p=1.00	p=1.00	p=1.00
Stationary noise -5 dB	p=1.00	p=0.11	p=0.17	p<0.0001	p<0.0001	p=0.001
Stationary noise 0 dB	p=0.20	p=1.00	p=1.00	p=0.56	p=0.33	p=1.00
Stationary noise +5 dB	p=0.005	p=0.96	p=0.96	p=1.00	p=0.98	p=1.00
Stationary noise +10 dB	p<0.0001	p=0.32	p=0.32	p=1.00	p=1.00	p=1.00
Stationary noise +15 dB	p<0.0001	p=0.33	p=0.32	p=1.00	p=1.00	p=1.00

The beta estimates for the performance scores and PPD across SNRs are shown in Table 3 in Appendix B.

The present study examined the interplay between hearing-status and two masker types on performance and PPD across a broad range of SNRs, covering low to high speech intelligibility levels for both NH and HI listeners. We hypothesized that the PPD would be smaller at low and high SNRs while a maximum PPD was expected in the mid-range of SNRs around 50% intelligibility. We furthermore hypothesized that NH listeners would show larger PPDs for difficult listening conditions compared to HI listeners and that the PPD changes would depend on the different masker types with generally larger PPDs for a single-talker masker

compared to a stationary noise masker (Koelewijn et al., 2014, 2012).
When the single-talker masker was present, maximum PPD was observed at the mid-range of SNRs while at lower and higher SNRs, smaller PPDs were measured for both groups of listeners. The PPD measured during sentence recognition in stationary noise showed a pronounced maximum PPD at -4 dB SNR for NH listeners, but a relatively small PPD for each of the SNRs resulted for HI listeners. The current study indeed demonstrated that the relation between PPD and SNR differs between NH and HI listeners, for both masker types. The results suggest that in HI as compared to NH listeners, the maximum PPD response is not only shifted but that the PPD is somewhat smaller (i.e., for both the stationary noise and interfering speech masker) but also varies less in response to SNR changes (stationary noise masker). We did not find larger PPDs for sentence recognition during the presence of a single-talker masker compared to the stationary noise masker. The results are discussed in detail below.

3.4 Discussion

Sentence recognition data

Sentence recognition performance was measured during the presence of a stationary noise and a single-talker masker across two separate SNR ranges. As expected and in line with (e. g. Festen and Plomp, 1990) an interactive effect of SNR and hearing ability (NH vs. HI) was observed in sentence recognition performance for both maskers (stationary noise, single-talker). Hearing impairment resulted in poorer sentence recognition performance, despite NAL-R correction except for the end points of the sentence recognition curves where both groups of listeners reached 0% or 100% correct sentence recognition. When both masker types were compared across the overlapping SNR range, the %correct sentence recognition functions for the stationary noise masker (slope NH = 13.7 %/dB; slope HI = 11.24 %/dB) were steeper than for the single-talker masker (slope NH = 5.6 %/dB; slope HI = 5.0 %/dB) at 50% correct performance. A difference between the performance curves of almost 10 dB indicates a benefit of the single-talker masker relative to the stationary noise for the NH listeners. The HI listeners had smaller fluctuating masker benefit as the estimated SNR for 50% correct (based on the fitted curves) only differed by 1 dB between masker types. The HI listeners also had a steeper curve for the stationary noise masker compared to the single-talker masker. In both groups there was a slight performance drop at +15 dB SNR for the single-talker masker, which was a surprise at such a high level of sentence audibility. Some listeners might have paid less attention at these relatively easy conditions. A single incorrectly repeated word could have resulted in a 5% lower performance as a sentence was only scored as correct if the listener correctly had repeated every word in the sentence.

Pupil data

The present study confirmed earlier findings showing that hearing loss influences the allocation of listening effort, as reflected by PPD, as function depending on the SNR (intelligibility). This was also shown by Kramer et al. (1997) and Zekveld et al. (2011) but with a limited range of SNR conditions and only for the stationary noise masker. The current study is the first to demonstrate that the PPD as function of SNR depends on the listener group (HI vs. NH). This is a highly interesting finding. We apply the FUEL framework (Pichora-Fuller et al., 2016) to interpret our findings. More specifically, we discuss how the expenditure of effort during listening may also depend on the listener's motivation to perform a task. We start with the discussion of the data for the stationary noise masker for the NH listeners.

When the stationary noise masker was presented, sentence recognition performance was very high across the range of positive SNRs from +16 dB to +8 dB. The corresponding PPD across these high SNRs was relatively low for the NH listeners, indicating low task demands. With further decreasing SNRs, sentence recognition performance in the NH group dropped from 100% at +4 dB to about 40% at -4 dB SNR. The corresponding PPD increased rapidly, especially between 0 dB and -4 dB SNR, where sentence recognition dropped abruptly. The NH listeners were probably motivated to keep up their high performance and they increased the amount of effort invested in the task with a maximum around -4 dB SNR. That most effort is invested at SNRs resulting in 50% performance levels may relate to the relatively steep psychometric function (%-correct) at this point. In this transition region, it pays to apply intense effort, which may drive the listener's motivation to keep on trying (Pichora-Fuller et al., 2016). When the SNR dropped from -4 dB to -12 dB, sentence recognition performance rapidly decreased to 0% correct. The corresponding PPD peaked over a relatively narrow SNR range and rapidly reduced when the task transitioned from difficult to impossible (as reflected in %-correct). This suggests that the listener's motivation may decrease as the application of intense effort brings no further reward in terms of maintaining sentence recognition performance (Pichora-Fuller et al., 2016). As such, these data contribute to the FUEL framework.

A key finding of this study is that HI listeners seem to be less adaptive in response to varying SNRs: they showed relatively small differences in PPD across the range of SNRs. The reduced PPD at +12 dB for the stationary masker (see Figure 2) might indicate release from masking, but as PPD increases again at +16 dB, it is more probably a random variation. Overall, fairly constant PPDs were shown for the HI listeners for the SNR range from +8 dB to -8 dB even though sentence recognition performance dropped from 100% correct at +8 dB SNR to 0% correct at -8 dB SNR. Given the nature of the energetic masker, interacting with the hearing impairment, the rewards of applying extra effort are probably relatively limited for HI listeners. On the other hand, the listener's peripheral impairment means that task demand is elevated even at high SNRs, when sentence recognition performance is high. The overall fairly constant PPDs across SNRs may indicate that when confronted with a stationary noise masker, HI listeners do not experience conditions providing much extra motivation for the expenditure of intense effort (Pichora-Fuller et al., 2016). Our second hypothesis that NH listeners would show larger PPD for difficult SNR conditions compared to HI listeners could not be confirmed for the stationary noise conditions. We did not find significantly different PPDs between groups for difficult SNR conditions.

The analysis revealed that PPD differs across SNRs within each group. For both groups of listeners, the PPD measured during sentence recognition at -4 dB and -8 dB SNR differed significantly from -12 dB SNR and for HI listeners PPDs differed even at high SNRs. Both groups perceive perhaps little motivation to apply intense effort at the SNR of -12 dB as it brings no reward in terms of improved sentence recognition performance. With increasing SNRs, PPDs increase and listeners of both groups may seem to realize that it pays to apply intense effort as their sentence recognition performance improves. Both groups of listeners seems to be able to spend intense effort but the NH listeners' motivation may decrease as the application of intense effort brings no further rewards in terms of maintaining sentence recognition performance at high SNRs while for HI listeners task demands are still elevated at high SNRs. It seems like the application of intense effort is shifted towards higher SNRs due to hearing-impairment but both groups of listeners are apparently able to invest intense effort.

Recently, Wu and colleagues (2016), used reaction times measured during different dual-task paradigms when speech recognition in stationary noise was the primary task. The results showed that the reaction time curves differed in shape between the young NH and the older HI listeners. Young NH listeners had longer reaction times during unfavorable SNRs and the reaction time curve was flatter than the curve of the older HI listeners. The older HI listeners had similar reaction times for both the unfavorable and favorable SNRs.

In the current study we found a significant interaction effect between SNR and listener group for the PPD measured during speech recognition in the stationary noise masker. However, we did not find a significant main effect of group. There was a trend that NH listeners had larger PPDs at unfavorable SNRs compared to HI listeners. Larger PPDs or longer reaction times measured in a dual-task paradigm scheme may indicate that NH listeners spent more effort during difficult listening situations compared to HI listeners. HI listeners on the other hand, showed a more flat PPD curve and similar reaction times, resulting from a dual-task paradigm, across unfavorable and favorable SNRs compared to NH listeners. The relatively flat pupil and reaction time functions may indicate that HI listeners are less sensitive across SNRs compared to NH listeners.

The pattern of results is slightly different for the single-talker masker. At high SNRs (+15 dB to +5 dB), where the NH listeners' recognition performance in the presence of the single-talker masker was close to 100%, PPDs were small. This may indicate low task demand. With further decreasing SNRs, sentence recognition performance in the NH group dropped from about 100% at +5 dB to about 55% at -10 dB SNR. The corresponding PPD increased and reached its maximum between -10 dB and -15 dB SNR, where sentence recognition dropped abruptly. At these SNRs, NH listeners seem still able to listen in dips of the competing speaker and segregate target and competing speech, which may motivate them to stay engaged in the task as sentence recognition performance is still between 25% and about 50% correct. Interestingly, the largest difference in PPD between NH and HI listeners was found when sentence recognition was measured at the SNR of -15 dB. NH listeners had significantly larger PPDs at -15 dB and at -20 dB SNR compared to HI listeners, which may support the assumption that NH listeners may be more engaged in the performance at -15 dB and at -20 dB SNR compared to HI listeners. At the very low SNRs, between -25 dB to -20 dB, PPDs for NH listeners were reduced which is probably due to giving up trying to perceive the speech as 0% correct sentence recognition performance was reached. This is in line with findings by

Zekveld & Kramer (2014) and the FUEL framework. An important finding of this study is that the PPD curve for the HI listeners was similar to that of the NH listeners but shifted about 10 dB upwards in SNR. Our first hypothesis, assuming smaller PPD at relatively low and high SNRs and maximum PPDs around 50% correct sentence recognition performance is true for both groups of listeners.

The intelligibility curve for the HI group was also parallel to the NH curve and shifted by about 10 dB. This suggests that the mechanisms in play are similar in both groups, and that the group difference lies in the matter of where along the SNR axis most effort is expended. We found significant interaction effects between group and SNRs for both masker types when the large SNR ranges were analyzed. When PPDs for both masker conditions (stationary and single-talker masker) were compared across the smaller SNR range (-10 to +15 dB SNR), the interaction effect between groups and SNRs was not significant. This is perhaps a consequence of the smaller SNR range, including less negative SNRs. When separate analyses were applied for each masker type, significant differences in PPD between listener groups were found for the low SNRs of -20 dB and -15 dB SNR for the single-talker masker. For the stationary noise masker, PPDs differed within each listener group with respect to the lowest SNR of -12 dB. The overlapping SNR range covered only SNRs from -10 dB to +15 dB and does not include those very low SNRs, where the interaction effects were found for the separate analyses. The hypothesized group differences at difficult SNRs could consequently not be confirmed for the SNR range between -10 dB and +15 dB. However, a significant interaction between SNR and masker type for the overlapping SNR range resulted. The significant interaction between SNR and masker type shows that the effect of masker type on the PPD depends on the SNR. The maximum PPDs for the separate SNR ranges of each masker type were obtained at SNRs very close to -5 dB (at -4 dB) for the NH group in stationary noise masker and at -10 dB SNR for both groups in single-talker masker. Surprisingly, PPDs measured during the presence of the single-talker masker were in general not significantly larger than PPDs measured for the stationary noise masker. This is contrary to previous findings, which showed larger PPDs for the presence of a single-talker masker compared to a fluctuating and a stationary noise masker (Koelewijn et al., 2014, 2012; Zekveld et al. 2014). Speech recognition in the presence of a single talker masker can result in a larger pupil response, which is likely due to additional informational masking which introduces more cognitive load than the presence of a stationary noise masker. This main effect of masker type on the PPD was previously shown for listeners of different age groups (e.g. young versus middle-aged listeners, Koelewijn et al., 2012) and with different hearing abilities (normal-hearing versus hearing-impaired, Koelewijn et al., 2014, 2012). In the current study, the sentence material was presented binaurally via headphones, in the same manner as in previous studies (Koelewijn et al., 2014, 2012; Zekveld et al. 2014). The main difference between the current study and previous studies lies in the implementation of different levels of intelligibility. In previous studies (Koelewijn et al., 2014, 2012; Zekveld et al. 2014), intelligibility was the independent (fixed) factor while in the current study fixed SNRs were tested. The differences in the experimental study design could perhaps affect differences in the pupil dilation for different masker types. Individually perceived differences of task difficulty and cognitive load might be larger across participants when SNRs are fixed compared to fixed levels of intelligibility. However, this is very speculative.

Overall, NH listeners had largest PPDs at negative SNRs while HI listeners had largest PPDs across a wide range of SNRs, when the single-talker masker was present during sentence recognition performance. Independent of the presented masker type, HI listeners had increased PPDs compared to NH listeners across a range of positive SNRs, which may indicate high task demands even though sentence recognition performance was high. If typical daily listening situations (SNRs) indeed tend to evoke elevated effort in HI but less in NH listeners, despite high performance levels in both groups, this could be a cause of the commonly reported fatigue in HI listeners (Hétu et al., 1988; Kramer et al., 2006). Further research is needed at this point, to conclude whether this is the case.

The aim of this study was to measure the PPD and % correct performance across the entire range of the psychometric function, including a large range of positive SNRs to cover real-life listening condition for HI listeners (Smeds et al., 2015; Lunner et al., 2016). We decided to present a range of fixed SNRs. Due to the large SNR range, we had relatively few SNRs in the mid-range of the psychometric function. A direct comparison of the PPD differences between groups and masker types based on these data is, however, not straightforward as the levels of the independent variable in this analysis (% correct sentence recognition) differ between groups and masker types. Also, there is a relatively small number of data points for the mid-range of sentence recognition performance. The present study demonstrates how the pupil response relates to changes in SNR. In previous studies (Zekveld & Kramer, 2014; Zekveld et al., 2011; Koelewijn et al., 2012), intelligibility was the independent (fixed) factor. The current and these previous studies indicate that both intelligibility and SNR influence the PPD – and that these effects cannot be differentiated across most of the psychometric intelligibility function where performance is not at floor or ceiling levels. However, SNR still influences the PPD at high intelligibility levels where performance is around 100% (current study), whereas the PPD can also differ between acoustically different conditions resulting in the same performance level (Koelewijn et al., 2012; Zekveld et al., 2014).

Our data demonstrate that pupillometry reveals effects that are not discovered using conventional speech in noise tests. Combined information about actual task demands and required effort during speech recognition performance is of great value for clinical applications such as measures of successful hearing aid fitting with respect to effortful listening in daily life environments (Ohlenforst et al., 2017).

Limitations

A possible limitation within this study may be the unequal number of participants with each group of listeners (NH and HI). We aimed to include 68 listeners in total and an equal numbers of NH and HI listeners, but unexpected drop outs and noisy pupil data limited our data collection. The analysis performed was selected as it provides relatively solid and reliable results while keeping the effect of the missing pupil data as small as possible. We did observe statistically significant interaction effects that replicate previous findings (Kramer et al., 2013; Zekveld et al., 2011). This supports the reliability of the present results.
A possible reason for the noisy pupil data may be the rather small number of sentences

(n=10 sentences) presented per SNR condition. The large number of SNRs (n=17 SNRs) across both masker conditions limited the amount of sentences we could present per condition (Versfeld et al., 2000). Previous studies applied adaptive speech reception threshold (SRT) procedures and used between 39 and 45 sentences for a reliable pupil dilation measurement (Zekveld et al., 2011; Koelewijn et al., 2012). The defined number of SNRs may be another limitation within this study. We chose a wide range of fixed SNRs to cover the whole range of sentence recognition performance. However, smaller steps (e.g. 2 dB steps) between those fixed SNRs may have provided an even clearer picture of the related effort, although we don't expect large deviations from the relatively gradual curves observed in the current study. The different types of background noise and the inclusion of NH versus HI listeners motivated us to choose different ranges of SNRs for the stationary and the single-talker masker. We wanted to make sure that even the NH listeners reached the lowest point of sentence recognition performance for both masker types. However, an equal range of SNRs for both masker types would have allowed a more direct comparison.

3.5 Conclusion

Our data indicate an interactive effect between hearing status (NH versus HI) and SNR on the PPD. The PPD changed depending on the difficulty of the listening condition and the listeners hearing abilities. With a stationary masker, NH listeners show maximum PPD across a relatively narrow range of low SNRs, while HI listeners show somewhat heightened PPD across a wide range of ecological SNRs. With a single-talker masker, NH and HI groups show similar PPD patterns, but with a shift towards higher SNRs in the HI listeners. These findings indicate that the pattern of listening and –more specifically – the allocation of effort during listening in daily life may be different for HI than for NH listeners. Further research is needed to find out if this is the reason why complaints of fatigue and stress are more often observed in HI than in NH listeners.

Acknowledgements

The authors wish to thank Artur Lorens from the Institute of Physiology and Pathology of Hearing in Warsaw for his thoughts and comments on this work. We would furthermore like to thank Elisabeth Wreford Andersen and Per Bruun Brockhoff from the Institute for Mathematics and Computer Science at the Technical University of Denmark (DTU Compute), for their support and advice with the statistical analyses. As well as Søren Laugesen and Niels Søgaard Jensen for their comments and input on the analysis. *This article presents*